Tracking and Recognition of Fist Move on Musical Excerpts at the Piano

Rastreamento e reconhecimento de movimentos de punho na execução de excertos musicais ao piano

Thyago P. Carvalho*†, Fabrizzio Alphonsus A. M. N. Soares*†, Carlos H. C. R. Costa *‡, Leandro Luís G. de Oliveira*†, Anderson da S. Soares*†, Luciana de O. Berretta*†, Ronaldo M. da Costa and*† Cristiane B. R. Ferreira*†

*Universidade Federal de Goias - UFG

†Instituto de informática - INF, Goiânia, Goiás

‡ Escola de Música e Artes Cênicas - EMAC, Goiânia, Goiás

***Abstract*—This works presents a method for teaching assistance in exercises of piano playing with hand tracking and gesture recognition. A tutor defines exercises to be performed and the system detects and tracks the hands of student and checks the correct perform of exercises. To hand detection and tracking are used coloured markers on the back of hands, captured by a regular webcam and for gesture recognition is used Multidimensional Dynamic Time Warping (MDTW), a n-dimensional version of Dynamic Time Warping (DTW). To verify the performance of the method, we have made videos of exercises, which were filtered for some adjusts and noise reduction. Results shows that the method has been capable to detect and recognize gestures successfully. Although the system needs some improvements, first impression about the system has been satisfactory. Thus, we hope, until the end of this work, implement a real-time version of the method for build a tutor system for piano classes.**

Keywords*—Piano teaching, Hand Tracking, Dynamic Time Warping

***Resumo*—Este trabalho apresenta um método para assistência ao ensino de excertos musicais ao piano com rastreio de mão e reconhecimento de gestos. Um tutor define os exercícios a serem realizados e o sistema detecta e rastreia as mãos do aluno e verifica a realização do excerto. Para detecção e rastreamento de mão são usados marcadores coloridos nas costas das mãos, capturada por uma Webcam regular e para reconhecimento de gestos é usado o *Multidimensional Dynamic Time Warping* (MDTW), uma versão n-dimensional do *Dynamic Time Warping* (DTW). Para verificar o desempenho do método, foram feitos vídeos de exercícios, que foram filtrados para alguns ajustes e redução de ruído. Os resultados mostram que o método foi capaz de detectar e reconhecer gestos com sucesso. Embora o sistema precise de algumas melhorias, a primeira impressão sobre o sistema tem sido satisfatória. Assim, esperamos que, até o final deste trabalho, sejamos capazes de implementar uma versão em tempo real do método para construir um sistema de auxílio as aulas de piano.**

Palavras-chaves*—Ensino de piano, rastreamento de mão, Dynamic Time Warping

I. Introdução

Conforme [1], o piano é um instrumento que o som pode ser produzido por um simples toque, sendo capaz de oferecer uma gama de variações de som que dependerá da forma como os braços são posicionados, a forma e a velocidade com que as teclas são tocadas. Ao executar um trecho de uma música, como um *legato*[1], existem inúmeras interpretações desta passagem, que podem variar de acordo com os movimentos dos dedos ou maior habilidade com o punho.

A correta utilização do punho dá maior destreza, suavidade, velocidade e conforto ao toque de teclas ao piano, o que consequentemente está diretamente ligado ao tom e controle de sonoridade [2].

A partitura é uma das documentações completas sobre uma obra musical a ser realizada, e que contém várias anotações que indicam notas musicais, suas alturas, duração, intervalos, a força a ser utilizada entre outros dados. No entanto, a partitura não contém dados que indicam como o trabalho deve ser interpretado. Assim, não há dados sobre os gestos a serem utilizados pelo pianista para caracterizar a interpretação e/o uso adequado dos músculos dos membros superiores. Embora seja extremamente importante usar os gestos durante a execução de obras no piano, não há documentação ou método científico que define como estes gestos devem ser usados.

O processo de ensino de um aluno de piano por um professor é complexo. O processo de ensino de piano é baseado em observação humana, pois não existe documentação e tudo é ensinado sem medições precisas ou dados de trajetória. A didática dos professores baseia-se em mostrar vídeos, pegar nas mãos dos estudantes para guiá-los e observar cuidadosamente a repetição dos exercícios.

O processo de ensino é altamente depende da habilidade do professor e, assim, muitas vezes o professor pode confundir-se e pensar que o aluno tenha aprendido os gestos corretamente. Portanto, uma vez que o aluno tenha se formado, ele/ela pode se tornar um professor que ensina gestos errados, ele/ela pode tocar músicas sem expressar o som esperado e, por fim, fazer movimentos errados e sofrer fadiga muscular, tendinite, entre outros problemas de saúde muscular.

Levando em consideração as questões explanadas temos o intuito de desenvolver um sistema, com reconhecimento de gestos das mãos em *excertos*[2] e obras musicais ao piano. O sistema deve observar, avaliar, e informar sobre os movimentos ao

[1] ligado, do Italiano, são notas que são executadas de forma ligada, ou seja a segunda nota é emitida antes da finalização do som da nota anterior.

[2] trecho retirado de uma obra musical; fragmento, passagem.

aluno permitindo que o aluno observe seu próprio movimento, receba dados de caminhos, velocidade, entre outros, durante a realização de um excerto.

O sistema irá desenvolver uma metodologia de ensino de piano de forma automatizada. Apoiando o aluno para que ele seja capaz de identificar os próprios movimentos do punho dependendo menos da observação do professor, mas de forma alguma reduzindo a necessidade do mesmo no processo.

II. Dados

Para a realização da experimentação do método foram gravados vídeos, a explicação dos materiais e métodos utilizados para gravação dos vídeos é feita na II-A. Antes da aplicação do método propriamente dito foi necessário realizar o pré-processamento dos vídeos, esta etapa é explicada na II-B.

A. Vídeos

Para a gravação dos vídeos foi utilizado uma câmera Olympus SP-810 UZ 14 megapixel com a resolução de $WXGA/720p$, taxa de 30 quadros por segundo e qualidade normal da imagem. A câmera foi posicionada com o auxílio de um tripé a 45 centímetros de altura da calda do piano. Após ser feito o posicionamento da câmera foi feita a gravação de 5 vídeos de cada um dos excertos propostos.

Para cada um dos excertos foi gravado 5 vídeos diferentes do mesmo. Os vídeos são classificados da seguinte forma:

1) Três vídeos com a execução correta. Exercícios esses que são utilizados para o treino do método
2) Três vídeos com erros inseridos aleatoriamente ao longo da execução do excerto.
3) Um vídeo com a execução correta, sem a movimentação do punho do pianista.

Para todos os vídeos com erros inseridos durante sua execução, foi feito uma análise empírica para detectar em que momento da execução o erro aconteceu, identificando o quadro inicial do erro e o quadro final do erro.

B. Pre-processamento

Apos a gravação de todos os vídeos, foi feito um pré-processamento externo ao algoritmo, cortando os vídeos para que eles iniciem e terminem de maneira mais próxima possível. A taxa de quadros considerados no processamento é de 5 por segundo ao invés dos 30 por segundo que a câmera utilizada para a gravação prove.

Também foi feito uma redução em uma escala de $escala\ original * \frac{2}{3}$ fazendo com que a resolução de 1280x720, obtida com a câmera, fique igual a 853x480. Lembrando que foi utilizado arredondamento no processo de redução pois está se trabalhando com *pixels*, e não existem quadros não inteiros.

III. Método proposto

O método proposto consiste no reconhecimento dos gestos da mão do pianista, para isso é necessário a detecção da mão, identificando quadro a quadro a direção da mão, e construir uma sequência temporal com o rastreamento da mão, a partir disso e necessário fazer a comparação destas séries com séries conhecidas para saber se o gesto executado está correto ou não.

Inicialmente, é necessário identificar o centro de massa da mão do pianista, para isso, foi inserido um marcador nesse membro. Posteriormente foi feito o rastreio do marcador para guardar esse dado para cada um dos *frames* selecionados do vídeo. As duas fases citadas são explicadas na Subseção III-A.

Com o rastreio do marcador adicionado ao membro do pianista, são gerados dados utilizados para mapear os movimentos da(s) mão(s) do pianista. Com uma base de dados pronta, foi utilizado um algoritmo de visão computacional para realizar o reconhecimento dos gestos capturados. Está faze é explicada na Subseção III-C.

Para se chegar a classificação se um bloco está ou não correto foi feito os seguintes passos:

1) Aplicou-se o MDTW entre todos os vídeos corretos, fazendo com que cada vídeo correto fosse comparado com os outros dois.
2) Obteve-se 3 distâncias para cada um dos blocos dos vídeos corretos.
3) Das 3 distâncias foi feito a média aritmética simples, obtendo assim um vetor médio de distâncias.
4) O vídeo é inserido e aplicado o MDTW com o melhor dos 3 vídeos corretos.
5) Então tem-se o vetor médio de distâncias corretas e o vetor de distâncias analisados
6) O resultado é determinado errado se sua distância for menor que o seu correspondente no vetor de distância média, e é determinado correto se sua distância for maior.

A. Detecção e Rastreamento

Como não é intenção fazer a manipulação do ambiente para mantê-lo controlado, foi tomado como referência [3], que utilizou marcadores para segmentar a bola de basebol e a mão do jogador. No presente trabalho foi utilizado um único marcador de cor predominantemente vermelha e forma retangular nas costas da mão do pianista, para realizar a correta identificação do centro de massa da mão.

Diferentemente do trabalho de [3], não é necessário a identificação dos diferentes lados para cada um dos diferentes objetos, por isso, existe a necessidade de apenas um marcador nas costas da mão do pianista. Com a inserção de um marcador retangular de $1x2$ cm de cor vermelha no ambiente é possível identificar o mesmo fazendo uma seleção nos quadros, lembrando que os mesmos estão no sistema de cor RGB, de $r = [142, 168]$, $g = [9, 26]$ e $b = [0, 19]$. A fig. 1 mostra as costas da mão do pianista com a presença do marcador.

Apos a identificação e isolamento apenas dos *pixels* refentes a área do marcador inserido, é possível calcular o centro do marcador representado pela Equação 1. O valor obtido é o centro de massa da mão do pianista.

$$BB_c = \frac{BB'' + BB'}{2} \tag{1}$$

Onde (BB_c) é o ponto central, considerado neste trabalho como o centro de massa da mão do pianista, do marcador

Fig. 1. Exemplo da mão do pianista com marcador

composto pelos pontos BB' e BB'' que são o canto superior esquerdo e o canto inferior direito, respectivamente.

Com a identificação do centro de massa da mão do pianista em cada *frame* a partir do método explicado acima. Monta-se uma série temporal, onde em cada momento temporal é armazenado os dados da posição do centro de massa da mão do pianista. Possibilitando assim, fazer o rastreamento da movimentação da(s) mão(s).

B. Pre-processamento

Ao fazer a comparação dos dados obtidos a partir dos vídeos gravados, por mais idênticos que sejam os gestos, eles acabam tendo valores de x e y diferentes. Fazer a normalização dos dados obtidos pelo rastreamento, faz com que eles se tornem mais próximos. Para se fazer a normalização das duas séries temporais com a função de normalização pelo desvio padrão Equação 2, também chamada de *Z-Score* [4]. Ao se aplicar a Equação 2 os valores da sequência temporal ficam normalizados de forma que a média desta sequência se torna zero e o desvio padrão um.

$$z = \frac{r_i - m_R}{dp_R} \tag{2}$$

Onde z é o resultado do elemento r_i normalizado, da sequência temporal R, sendo m_R e dp_R a média e o desvio padrão da sequência temporal R, respectivamente.

C. Reconhecimento de gesto

Para o reconhecimento dos gestos é utilizado um algoritmo de visão computacional, possibilitando a realização de comparações entre séries temporais distintas. É possível identificar um gesto inserido no sistema comparando com outros gestos já conhecidos

1) DTW: O *Dynamic time warping* (DTW) que é citado em diversos artigos [5], [6], [7] e [8] teve sua aplicação, para diversas áreas e para vários problemas, intensificada posteriormente para desenvolvimento de novas soluções. Apesar deste método ter uma complexidade de $O\left(n^2\right)$ ele é um dos melhores recursos para uma grande variedade de domínios por exemplo a bioinformática, medicina, engenharia, entretenimento e etc.

O DTW é uma técnica que alinha duas sequências de dados semelhantes que diferem em seu tamanho. É um algoritmo que tem como objetivo encontrar o melhor alinhamento entre duas sequências unidimensionais e descobrir a distância entre elas. A fig. 2 foi alterada apos ter sido retirada do trabalho de [9],

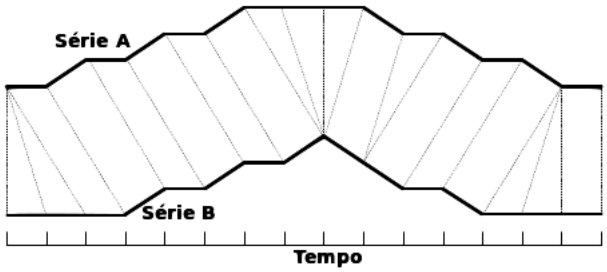

Fig. 2. Exemplo de alinhamento de duas séries temporais

e demonstra como é feito o alinhamento séries temporais (A e B) através da aplicação do DTW.

Quando se considera duas sequências unidimensionais $R = r_1, r_2, ..., r_N$ e $T = t_1, t_2, ..., t_M$ de tamanhos N e M respectivamente, e que uma sequência não é linearmente maior ou menor do que a outra no eixo do tempo, fazendo que consequentemente $N \neq M$.

O primeiro passo do Algoritmo do DTW é calcular a distância entre cada um dos elementos de cada uma das séries temporais, o cálculo é feito conforme a Equação 3.

$$d(i,j) = (R(i) - T(j))^2 \tag{3}$$

Onde $R(i)$ é o vetor temporal R na posição i, $T(j)$ é o vetor temporal T na posição j, com o quadrado da diferença desses valores consegue-se calcular o valor de $d(i,j)$ sendo d a matriz de distância na linha i e coluna j.

O segundo passo do Algoritmo do DTW é construir a matriz de distância acumulada conforme a Equação 4,que é construída com base no algoritmo de programação dinâmica. Com essa expressão é possível obter o custo de associação de $R(i)$ e $T(j)$.

$$D(i,j) = d(i,j) + min \begin{cases} D(i-1,j) \\ D(i,j-1) \\ D(i-1,j-1) \end{cases} \tag{4}$$

Onde $D(i,j)$ é a matriz de distância acumulada na linha i e coluna j com o valor calculado utilizando a soma da matriz de distância $d(i,j)$ explicada na Equação 3 somado com o menor valor das três posições vizinhas anteriores à posição referida.

Como existem problemas que tem a necessidade de trabalhar com sequências temporais multidimensionais, o DTW já não consegue mais resolver esse escopo de problemas. Conforme [8] quando existe a necessidade de se alinhar sequências multidimensionais, a maneira mais simples é aplicar o DTW em cada uma das dimensões das séries temporais. Fazendo com que a soma das distâncias de todas as dimensões possa mostrar a semelhança entre duas sequências. Entretanto, os pontos correlacionados pela sincronização não estarão corretos. De forma a gerar uma solução para essa brecha presente no DTW e deixa-lo mais genérico foi criado o MDTW (Multipledimensional Dynamic Time Warping) que tem seu funcionamento explanado na Subseção III-C2.

2) MDTW: Considerando então duas sequências multidimensionais $R \in R^{K*M}$ e $T \in R^{K*N}$, onde K é o numero de dimensões e M e N são os respectivos tamanhos das sequências.

O primeiro passo do algoritimo MDTW e executar o cálculo da matriz de distância conforme Equação 5, que é bem semelhante a Equação 3 que foi apresentada anteriormente, com o acréscimo do cálculo da distância de sequências temporais multidimensionais e não somente unidimensionais. Com essa expressão é possível calcular o somatório da distância entre cada um dos elementos das séries temporais multidimensionais.

$$d(i,j) = \sum_{k=1}^{K} \left(R(i,k) - T(j,k)\right))^2 \quad (5)$$

Onde $R(i,k)$ é o vetor temporal R na dimensão k e posição i, $T(i,k)$ é o vetor temporal T na dimensão k e posição j, com o somatório do quadrado da diferença desses valores em cada dimensão k consegue-se calcular o valor de $d(i,j)$ sendo d é a matriz de distância na linha i e coluna j.

O segundo passo do algoritimo é construir a matriz de distância acumulada, que é feito igualmente ao algoritmo DTW explicado na Subseção III-C1, utilizando a Equação 4.

IV. Análise dos resultados

A Table I. mostra a matriz de confusão construída a partir dos resultados obtidos com aplicação do método sobre os dados, a matriz confusão é um método bastante conhecido para tornar mais intuitivo os resultados da qualidade de uma predição.

TABLE I. Matriz de confusão

Class	P	N
Y	152.00	13.00
N	33.00	34.00
Total	185.00	47.00

Na Table I. o que interessa positivamente é a diagonal primaria, que mostram onde o classificador conseguiu acertar na predição de certo ou errado. Mas não pode-se deixar de lado a diagonal secundária que tem resultados interessantes a serem considerados mostrando o quanto o classificador está fazendo predições erronias, e qual o tipo de predição erronia tem maior valor.

Uma observação sobre um dos valores da Table I. que deve ser ressaltada é que o valor de falso negativo (FN) é 33. O fato do valor de FN ser muito maior que o valor de falso positivo, mostra que o método é muito rigoroso.

A partir da Table I. é possível calcular algumas métricas que auxiliam na compreensão dos dados. As métricas calculadas são apresentadas na Table II. Os valores apresentados na Tabela II. mostram que o método conseguiu valores altos na precisão e acurácia. Demonstrando que tende a ser um método bom, mesmo que ainda precise de melhorias tanto para não ser excessivamente rigoroso nas predições e com uma tacha de acerto um pouco maior.

TABLE II. Tabela de métricas obtidas através da tabela confusão

Métricas	Valores
Taxa de FP	0.28
Taxa de TP	0.82
Precisão	0.92
Acurácia	0.80
Medida-F	0.87

V. Conclusão

Este trabalho propôs a implementação de um sistema para auxiliar o aprendizado de movimentos ao piano com fins interpretativos. O principal objetivo é apoiar o ensino, permitindo o aluno comparar seus próprios movimentos com exemplos propostos pelo professor. O sistema, embora ainda em desenvolvimento, nos experimentos realizados, apresentou um desempenho satisfatório no rastreamento e reconhecimento de gestos. Assim, podemos concluir que o sistema apresenta resultados promissores e poderá no futuro ser utilizado como ferramenta de auxílio ao ensino.

VI. Trabalhos Futuros

Pretende-se ainda desenvolver uma interface interativa para que os alunos possam visualizar-se realizando os exercícios em tempo real. Desta forma, obter informações e sugestões sobre os movimentos executados ao piano. Pretende-se também, realizar avaliação de usabilidade do sistema por alunos, vizando avalizar sua real empregabilidade na tarefa de ensino/aprendizagem.

VII. Agradecimentos

Este trabalho foi realizado com apoio financeiro dos editais 5, 6 e 8/2012 da Fundação de Apoio à Pesquisa do Estado de Goiás (FAPEG).

References

[1] Azako Tamura, *A arte pianística de Magda Tagliaferro*, 1st ed. SP: Fundação Magda Tagliaferro, 1990, tradução de Dirce Kimyo Miyamura, 1997.

[2] Tarcísio Gomes Filho, "O legado pedagógico de isabelle vengerova: Um estudo de aplicação do conceitos sobre técnica pianística," Ph.D. dissertation, Programa de Pós graduação em Música na Unicamp, Campinas, SP, 2008.

[3] C. Theobalt, I. Albrecht, J. Haber, M. Magnor, and H.-P. Seidel, "Pitching a baseball: tracking high-speed motion with multi-exposure images," in *ACM Transactions on Graphics (TOG)*, vol. 23. ACM, 2004, pp. 540–547.

[4] Neil J. Salkind, *Encyclopedia of Research Design*. SAGE Publications, Inc., 2010.

[5] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[6] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *Third Workshop on Mining Temporal and Sequential Data*, 2004, pp. 22–25.

[7] G. A. Ten Holt, M. J. T. Reinders, and E. A. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Thirteenth annual conference of the Advanced School for Computing and Imaging*, vol. 300, 2007.

[8] P. Sanguansat, "Multiple multidimensional sequence alignment using generalized dynamic time warping," *WSEAS Transaction on Mathematics*, vol. 11, no. 8, pp. 684–694, 2012.

[9] Salvadore S. and Chan P., "FastDTW toward accurate dynamic time warping in linear time and space," 2004.